Une version discriminante de l'Indice Probabiliste d'Ecart à l'Equilibre pour mesurer la qualité des règles

Julien Blanchard, Fabrice Guillet
Henri Briand, Régis Gras

LINA – FRE 2729 CNRS
Polytech'Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Résumé. L'évaluation de la qualité de connaissances est une étape clef dans un processus de découverte de règles d'association. Afin de mesurer la significativité statistique de l'écart à l'équilibre, nous avons proposé dans (Blanchard *et al.* 2005,) un nouvel indice de qualité nommé *IPEE*, fondé sur un modèle probabiliste. En tant que mesure statistique, *IPEE* a le désavantage d'être peu discriminant quand les cardinaux des données sont grands (de l'ordre de 10^4). Dans cet article, nous proposons d'adapter *IPEE* à ce type de données en reprenant le principe de l'intensité d'implication entropique, c'est-à-dire en associant *IPEE* avec une mesure entropique.

1 Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d'association (Agrawal *et al.* 1993,) sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives $a \rightarrow b$ où a et b sont des conjonctions d'items (variables booléennes de la forme $attribut = valeur$). Une telle règle signifie que la plupart des enregistrements qui vérifient la prémisse a dans la base de données vérifient aussi la conclusion b.

Une étape cruciale dans un processus de découverte de règles d'association est la validation des règles après leur extraction. En effet, de par leur nature non supervisée, les algorithmes de data mining peuvent produire des règles en très grande quantité et dont beaucoup sont sans intérêt. Pour aider le décideur (expert des données étudiées) à trouver des connaissances pertinentes parmi ces résultats, l'une des principales solutions consiste à évaluer et ordonner les règles par des mesures de qualité. Il en existe deux catégories : les subjectives (orientées décideur) et les objectives (orientées données). Les mesures subjectives prennent en compte les objectifs du décideur et ses connaissances *a priori* sur les données (Liu *et al.* 2000,) (Padmanabhan & Tuzhilin 1999,) (Silberschatz & Tuzhilin 1996,). En revanche, seuls les cardinaux liés à la contingence des données interviennent dans le calcul des mesures objectives (Tan *et al.* 2004,) (Guillet 2004,) (Lallich & Teytaud 2004,) (Lenca *et al.* 2004,).

Dans (Blanchard *et al.* 2005,), nous avons introduit une nouvelle mesure objective de qualité de règle : l'*Indice Probabiliste d'Ecart à l'Equilibre* (*IPEE*). Par *équilibre*, nous désignons l'incertitude maximale de la conclusion sachant que la prémisse est vraie, c'est-à-dire le fait qu'une règle possède autant de contre-exemples que d'exemples. *IPEE* évalue la significativité statistique de l'écart à l'équilibre au regard d'un modèle probabiliste (là où l'intensité d'implication (Gras 1996,) ou l'indice de vraisemblance du lien (Lerman 1981,) par exemple évaluent la significativité statistique de l'écart à l'indépendance). Cette nouvelle mesure est le seul indice d'écart à l'équilibre qui soit de nature statistique (voir tableau 1, et (Guillet 2004,) pour les références bibliographiques). La nature statistique fait à la fois la force et la faiblesse de *IPEE* (Blanchard *et al.* 2005,) :

– L'indice prend en compte la taille des phénomènes étudiés. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données.

	Indice d'écart à l'équilibre	**Indice d'écart à l'indépendance**
Indice descriptif	– confiance, – indice de Sebag et Schoenauer, – taux des exemples et contre-exemples, – indice de Ganascia, – moindre-contradiction, – indice d'inclusion...	– coefficient de corrélation, – indice de Loevinger, – lift, – conviction, – TIC, – rapport de cote, – multiplicateur de cote...
Indice statistique	IPEE	– intensité d'implication, – indice d'implication, – indice de vraisemblance du lien, – contribution orientée au χ^2, – rule-interest...

TAB. 1 – Classification des mesures objectives de qualité de règle

– Cependant, l'indice devient peu discriminant[1] quand la taille des phénomènes étudiés est grande (de l'ordre de 10^4). Ceci s'explique par le fait que même des écarts triviaux peuvent, au regard d'effectifs importants, s'avérer statistiquement significatifs.

Dans cet article, nous proposons d'adapter *IPEE* aux jeux de données de grande taille. Reprenant le principe de l'intensité d'implication entropique (GRAS *et al.* 2001*b*,) (GRAS *et al.* 2001*a*,) (BLANCHARD *et al.* 2004,), l'adaptation consiste à associer *IPEE* à une mesure entropique de qualité de règle, qui elle n'est pas de nature statistique, l'*indice d'inclusion*. Dans la partie suivante, nous rappelons les fondements de *IPEE* et de l'indice d'inclusion. Puis nous présentons la version discriminante de *IPEE*, nommée *IP3E* pour *IPEE Entropique*, et étudions ses propriétés.

2 Rappels sur *IPEE* et l'indice d'inclusion

Nous considérons un ensemble E de n objets décrits par des variables booléennes. Dans le vocabulaire des règles d'association, les objets sont des transactions enregistrées dans une base de données, les variables sont appelées des items, et les conjonctions de variables des itemsets. Etant donné un itemset a, nous notons A l'ensemble des transactions qui vérifient a, et n_a le cardinal de A. Le complémentaire de A dans E est l'ensemble $\overline{A}$ de cardinal $n_{\overline{a}}$. Une règle d'association est un couple (a, b) noté $a \rightarrow b$ où a et b sont deux itemsets qui ne possèdent pas d'item en commun. Les exemples de la règle sont les objets qui vérifient la prémisse a et la conclusion b (objets de $A \cap B$), tandis que les contre-exemples sont les objets qui vérifient a mais pas b (objets de $A \cap \overline{B}$). Dans la suite, nous appelons "variables" les itemsets.

2.1 *IPEE*

Etant donnée une règle $a \rightarrow b$, *IPEE* mesure la significativité statistique de l'écart à l'équilibre de la règle. La configuration d'équilibre étant définie par l'équirépartition dans A des exemples $A \cap B$ et des contre-exemples $A \cap \overline{B}$, l'hypothèse de référence est l'hypothèse H_0 d'équiprobabilité entre les exemples et les contre-exemples. Associons donc à l'ensemble A un ensemble aléatoire X de cardinal n_a tiré dans E sous cette hypothèse : $\mathrm{P}(X \cap B) = \mathrm{P}(X \cap \overline{B})$ (voir figure 1). Le nombre de contre-exemples attendu sous H_0 est le cardinal de $X \cap \overline{B}$, noté $\left|X \cap \overline{B}\right|$. Il s'agit d'une variable aléatoire dont $n_{a\overline{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

Définition 1 L'**indice probabiliste d'écart à l'équilibre (*IPEE*)** d'une règle $a \rightarrow b$ est défini par :

$$IPEE(a \rightarrow b) = \mathrm{P}(\left|X \cap \overline{B}\right| > n_{a\overline{b}} \mid \mathrm{H}_0)$$

Une règle $a \rightarrow b$ est dite admissible au seuil de confiance $1 - \alpha$ si $IPEE(a \rightarrow b) \geq 1 - \alpha$.

[1] Les règles sont jugées soit très bonnes (valeurs proches de 1), soit très mauvaises (valeurs proches de 0).

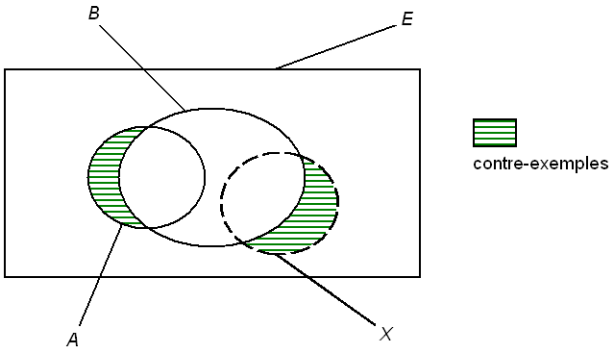


FIG. 1 – Tirage aléatoire d'un ensemble X sous hypothèse d'équiprobabilité entre les exemples et les contre-exemples

IPEE quantifie donc l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ eu égard à l'hypothèse H_0. En particulier, si $IPEE(a \rightarrow b)$ est proche de 1 alors il est invraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables. Cet indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse (et α comme le risque de première espèce de ce test). Toutefois, à l'instar de l'intensité d'implication et de l'indice de vraisemblance du lien (où H_0 est l'hypothèse d'indépendance entre a et b), il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

Dans le cadre d'un tirage avec remise, $\left|X \cap \overline{B}\right|$ suit une loi binomiale de paramètres $\frac{1}{2}$ (autant de chances de tirer un exemple que de tirer un contre-exemple) et n_a. *IPEE* s'écrit donc :

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} \mathrm{C}_{n_a}^{k}$$

Quand $n_a \geq 20$, la loi binomiale peut être approximée par la loi normale de moyenne $\frac{n_a}{2}$ et d'écart-type $\sqrt{\frac{n_a}{4}}$. *IPEE* ne dépend ni de n_b, ni de n puisque l'hypothèse d'équilibre H_0 ne se définit pas à l'aide de n_b et de n (contrairement à l'hypothèse d'indépendance).

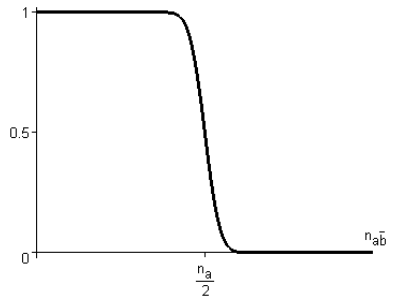


FIG. 2 – Représentation de *IPEE* en fonction de $n_{a\bar{b}}$

2.2 Indice d'inclusion

Gras a fondé l'indice d'inclusion sur deux écarts à l'équilibre :

– l'écart à l'équilibre de la règle à évaluer $a \rightarrow b$, associé au déséquilibre entre les exemples $A \cap B$ et les contre-exemples $A \cap \overline{B}$,

– l'écart à l'équilibre de la règle contraposée $\overline{b} \rightarrow \overline{a}$,associé au déséquilibre entre les exemples $\overline{A} \cap \overline{B}$ et les contre-exemples $A \cap \overline{B}$.

Une mesure bien connue pour évaluer les déséquilibres de façon non linéaire est l'entropie de Shannon (**?**). Considérons l'expérience aléatoire qui consiste à vérifier si b est vrai quand a est vrai. L'incertitude moyenne de l'expérience est donnée par l'entropie conditionnelle $H(b/a=1)$ de la variable b sachant la réalisation de a [2] :

$$H(b/a=1) = -\frac{n_{ab}}{n_a}\log_2\frac{n_{ab}}{n_a} - \frac{n_{a\overline{b}}}{n_a}\log_2\frac{n_{a\overline{b}}}{n_a}$$

Similairement, l'entropie conditionnelle $H(\overline{a}/b=0)$ quantifie l'incertitude moyenne de l'expérience aléatoire qui consiste à vérifier si a est faux quand b est faux :

$$H(\overline{a}/b=0) = -\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}}\log_2\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} - \frac{n_{a\overline{b}}}{n_{\overline{b}}}\log_2\frac{n_{a\overline{b}}}{n_{\overline{b}}}$$

L'indice d'inclusion utilise les entropies conditionnelles $H(b/a=1)$ et $H(\overline{a}/b=0)$ pour mesurer les écarts à l'équilibre d'une règle et de sa contraposée. Des règles de bonne qualité du point de vue de l'écart à l'équilibre engendrent des entropies faibles. Pour obtenir une mesure unique, les entropies sont combinées par la moyenne géométrique :

$$\sqrt{(1-H(b/a=1))(1-H(\overline{a}/b=0))}$$

Les compléments à 1 permettent d'associer les règles de bonne qualité aux valeurs fortes et non aux valeurs faibles. Pour renforcer le contraste entre les petites et les grandes entropies, celles-ci sont élevées à la puissance d'un nombre réel fixé $\omega \geq 1$:

$$\sqrt[2\omega]{(1-H(b/a=1)\ ^{\omega})(1-H(\overline{a}/b=0)\ ^{\omega})}$$

Par leur symétrie, les entropies conditionnelles $H(b/a=1)$ et $H(\overline{a}/b=0)$ évaluent identiquement un déséquilibre en faveur des exemples et un déséquilibre en faveur des contre-exemples : $H(b/a=1) = H(\overline{b}/a=1)$ et $H(\overline{a}/b=0) = H(a/b=0)$. Afin d'obtenir une mesure de qualité de règle, seul le déséquilibre en faveur des exemples doit être retenu. Pour cela, comme dans (Gras *et al.* 2001*a*,) et (Blanchard *et al.* 2004,), nous considérons qu'une règle est sans intérêt lorsque les contre-exemples sont plus nombreux que les exemples, et annulons les termes $1-H(b/a=1)^{\ \omega}$ et $1-H(\overline{a}/b=0)^{\ \omega}$ quand $n_{a\overline{b}} \geq \frac{n_a}{2}$ et $n_{a\overline{b}} \geq \frac{n_{\overline{b}}}{2}$ respectivement. Une autre solution est proposée dans (Gras *et al.* 2001*b*,) mais elle est moins intuitive et engendre une mesure de qualité moins filtrante.

Définition 2 L'**indice d'inclusion** τ d'une règle $a \rightarrow b$ est défini par :

$$\tau(a \rightarrow b) = \sqrt[2\omega]{I^{\omega}_{b/a=1} I^{\omega}_{\overline{a}/b=0}} \quad \text{où } \omega \geq 1 \text{ et :}$$

$$I^{\omega}_{b/a=1} = \begin{cases} 1-\left(-\frac{n_{ab}}{n_a}\log_2\frac{n_{ab}}{n_a} - \frac{n_{a\overline{b}}}{n_a}\log_2\frac{n_{a\overline{b}}}{n_a}\right)^{\omega} & \text{si } n_{a\overline{b}} < \frac{n_a}{2} \\ 0 & \text{sinon} \end{cases}$$

$$I^{\omega}_{\overline{a}/b=0} = \begin{cases} 1-\left(-\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}}\log_2\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} - \frac{n_{a\overline{b}}}{n_{\overline{b}}}\log_2\frac{n_{a\overline{b}}}{n_{\overline{b}}}\right)^{\omega} & \text{si } n_{a\overline{b}} < \frac{n_{\overline{b}}}{2} \\ 0 & \text{sinon} \end{cases}$$

L'indice d'inclusion s'annule dès que l'écart à l'équilibre de la règle ou de sa contraposée n'est pas orienté en faveur des exemples, c'est-à-dire lorsque $n_{a\overline{b}} \geq \min(\frac{n_a}{2}, \frac{n_{\overline{b}}}{2})$.

[2] Les fonctions entropiques associent des variables et des réalisations de variables. Pour plus de clarté, les réalisations d'une variable booléenne a sont notées $a=1$ et $a=0$ dans les fonctions entropiques, et non pas a et $\overline{a}$ comme dans les autres notations.

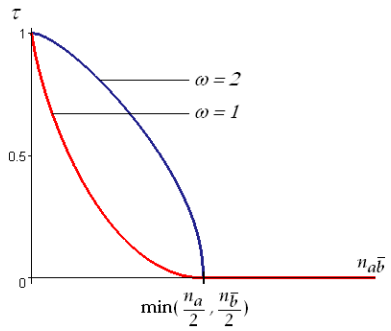


FIG. 3 – Représentation de l'indice d'inclusion en fonction de $n_{a\bar{b}}$

ω est un paramètre de sélectivité de l'indice d'inclusion qui peut être ajusté en fonction des données étudiées : plus ω est faible, plus l'indice d'inclusion décroît rapidement avec les contre-exemples, et plus le filtrage des règles est sévère (figure 3). En analyse statistique implicative, c'est généralement $\omega = 2$ qui est retenu. Ce choix permet que l'indice d'inclusion réagisse faiblement aux premiers contre-exemples (un faible nombre de contre-exemples ne doit pas remettre en cause la règle). Nous choisissons dans la suite $\omega = 2$.

3 La mesure *IP3E*

3.1 Définition

Comme dans l'intensité d'implication entropique (GRAS *et al.* 2001*b*,) (GRAS *et al.* 2001*a*,), nous associons *IPEE* et l'indice d'inclusion par la moyenne géométrique. Afin que l'indice d'inclusion vale 0.5 à l'équilibre comme *IPEE*, nous lui appliquons une transformation affine.

Définition 3 L'**Indice Probabiliste Entropique d'Ecart à l'Equilibre** (*IP3E*) d'une règle $a \rightarrow b$ est définie par :

$$IP3E(a \rightarrow b) = \sqrt{IPEE(a \rightarrow b) \times \frac{1}{2}(\tau(a \rightarrow b) + 1)}$$

3.2 Propriétés

La mesure *IP3E* prend ses valeurs dans $[0;1]$. Elle vaut $\sqrt{1 - \frac{1}{2^{n_a}}}$ pour une règle logique (sans contre-exemple), et 0.5 pour une règle à l'équilibre. *IP3E* est représenté en fonction du nombre de contre-exemples dans la figure 4. Nous pouvons voir que la mesure conserve les mêmes comportements intuitivement satisfaisants que *IPEE* :

- *IP3E* réagit faiblement aux premiers contre-exemples (décroissance lente).
- Le rejet des règles s'accélère dans une zone d'incertitude autour de l'équilibre $n_{a\bar{b}} = \frac{n_a}{2}$ (décroissance rapide).

Dans les figures 5.(a) et (b), les effectifs des données sont multipliés par un coefficient γ à partir d'une configuration initiale. Nous pouvons voir que les valeurs prises par *IPEE* et *IP3E* se rapprochent de 0 ou 1 quand les effectifs grandissent. Ceci illustre la nature statistique des deux mesures : plus la taille des phénomènes étudiés est grande, plus l'écart à l'équilibre observé dans les données est statistiquement

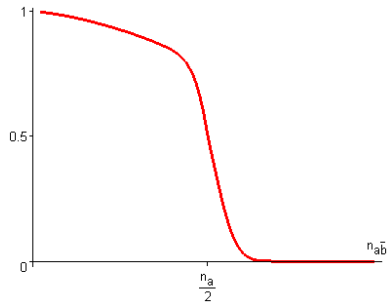
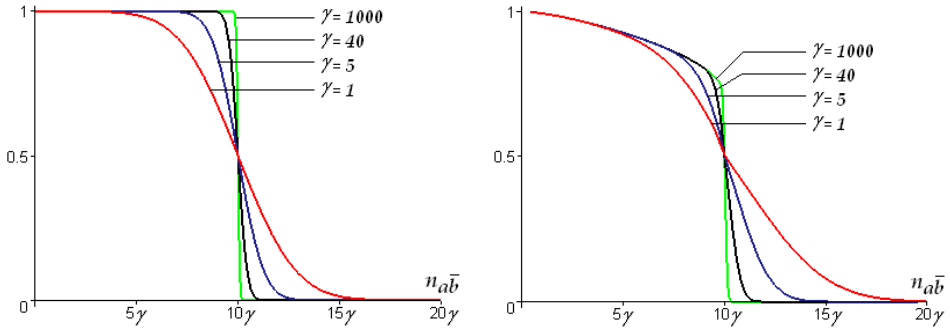


FIG. 4 – Représentation de *IP3E* en fonction de $n_{a\bar{b}}$



(a) *IPEE* (b) *IP3E*

FIG. 5 – Représentations de *IPEE* et *IP3E* avec la dilatation des effectifs
($n_a = 20 \times \gamma,\ n_b = 50 \times \gamma,\ n = 100 \times \gamma,$
$n_{a\bar{b}} \in [0 \times \gamma\ ;\ 20 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

significatif, et plus on peut confirmer la bonne ou la mauvaise qualité de la règle. Cependant, alors que *IPEE* devient peu discriminante (comportement "binaire") quand les effectifs étudiés grandissent (figure 5.(a)), *IP3E* a l'avantage de rester discriminante (5.(b)).

4 Conclusion

Dans cet article, nous avons présenté une mesure objective de qualité de règles qui évalue l'écart à l'équilibre. Il s'agit d'une version de *IPEE* adaptée aux jeux de données de grande taille selon le principe de l'intensité d'implication entropique : les valeurs de *IPEE* sont modulées par un indice descriptif fondé sur l'entropie de Shannon. La mesure qui en résulte, nommée *IP3E*, est de nature statistique mais reste discriminante quand la taille des phénomènes étudiés est grande. Elle peut être vue comme l'analogue de l'intensité d'implication entropique pour l'écart à l'équilibre.

Références

Rakesh AGRAWAL, Tomasz IMIELIENSKI & Arun SWAMI. «Mining association rules between sets of items in large databases». Dans : «Proceedings of the 1993 ACM SIGMOD international conference on management of data», ACM Press, 1993, pages 207–216.

Julien Blanchard, Fabrice Guillet, Henri Briand & Régis Gras. «*IPEE* : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles». Dans : «Actes de l'Atelier Qualité des Données et des Connaissances DKQ 2005, associé à EGC 2005», , 2005, pages 175–179.

Julien Blanchard, Pascale Kuntz, Fabrice Guillet & Régis Gras. «Mesure de la qualité des règles d'association par l'intensité d'implication entropique». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 33–43. Numéro spécial Mesures de qualité pour la fouille de données.

Régis Gras. *L'implication statistique : nouvelle méthode exploratoire de données.* La Pensée Sauvage Editions, 1996.

Régis Gras, Pascale Kuntz & Henri Briand. «Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données». Dans : *Mathématiques et Sciences Humaines*, tome 39 n° 154-155, 2001*a*, pages 9–29.

Régis Gras, Pascale Kuntz, Raphaël Couturier & Fabrice Guillet. «Une version entropique de l'intensité d'implication pour les corpus volumineux». Dans : *Extraction des Connaissances et Apprentissage*, tome 1 n° 1-2, 2001*b*, pages 69–80. Actes des journées Extraction et Gestion des Connaissances (EGC) 2001.

Fabrice Guillet. «Mesures de la qualité des connaissances en ECD», 2004. Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

Stéphane Lallich & Olivier Teytaud. «Evaluation et validation de l'intérêt des règles d'association». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 193–218. Numéro spécial Mesures de qualité pour la fouille de données.

Philippe Lenca, Patrick Meyer, Benoît Vaillant, Philippe Picouet & Stéphane Lallich. «Evaluation et analyse multicritère des mesures de qualité des règles d'association». Dans : *Revue des Nouvelles Technologies de l'Information*, tome E-1, 2004, pages 219–246. Numéro spécial Mesures de qualité pour la fouille de données.

Israël César Lerman. *Classification et analyse ordinale des données.* Dunod, 1981.

Bing Liu, Wynne Hsu, Shu Chen & Yiming Ma. «Analyzing the subjective interestingness of association rules». Dans : *IEEE Intelligent Systems*, tome 15 n° 5, 2000, pages 47–55.

Balaji Padmanabhan & Alexander Tuzhilin. «Unexpectedness as a measure of interestingness in knowledge discovery». Dans : *Decision Support Systems*, tome 27 n° 3, 1999, pages 303–318.

Avi Silberschatz & Alexander Tuzhilin. «What makes patterns interesting in knowledge discovery systems». Dans : *IEEE Transactions on Knowledge and Data Engineering*, tome 8 n° 6, 1996, pages 970–974.

Pang-Ning Tan, Vipin Kumar & Jaideep Srivastava. «Selecting the right objective measure for association analysis». Dans : *Information Systems*, tome 29 n° 4, 2004, pages 293–313.

Summary

Assessing rules with interestingness measures is the cornerstone of successful applications of association rule discovery. In order to evaluate the statistical significance of the deviation from equilibrium, we proposed in (Blanchard *et al.* 2005,) a new measure of interestingness named $IPEE$, based on a probabilistic model. Like the other statistical measures, the main drawback of $IPEE$ is that it has a low discriminating power when the cardinalities in the data are large. In this article, we propose to adapt $IPEE$ to this kind of datasets by following the entropic implication intensity principle, i.e. by combining $IPEE$ with an entropic measure.